

Introduction

Overview of tape archive system

- file systems in direct support of tape archives
- computer systems for computing and archiving
- the big picture
- focus on Data Migration Facility (DMF)
- what to do and why
- what not to do and why

Changes since 4/26/2012

- discussion about two tape copy requests
- **any archive file not *sitetagged* with "2" by May 31, 2012 is subject to have its second tape copy removed**

Main Systems

- Where do I compute?
 - Discover
 - Dali
- Where is the tape archive?
 - Physically closest to Dirac
 - Virtually
 - Everywhere (Discover/Dali)

Good to Know

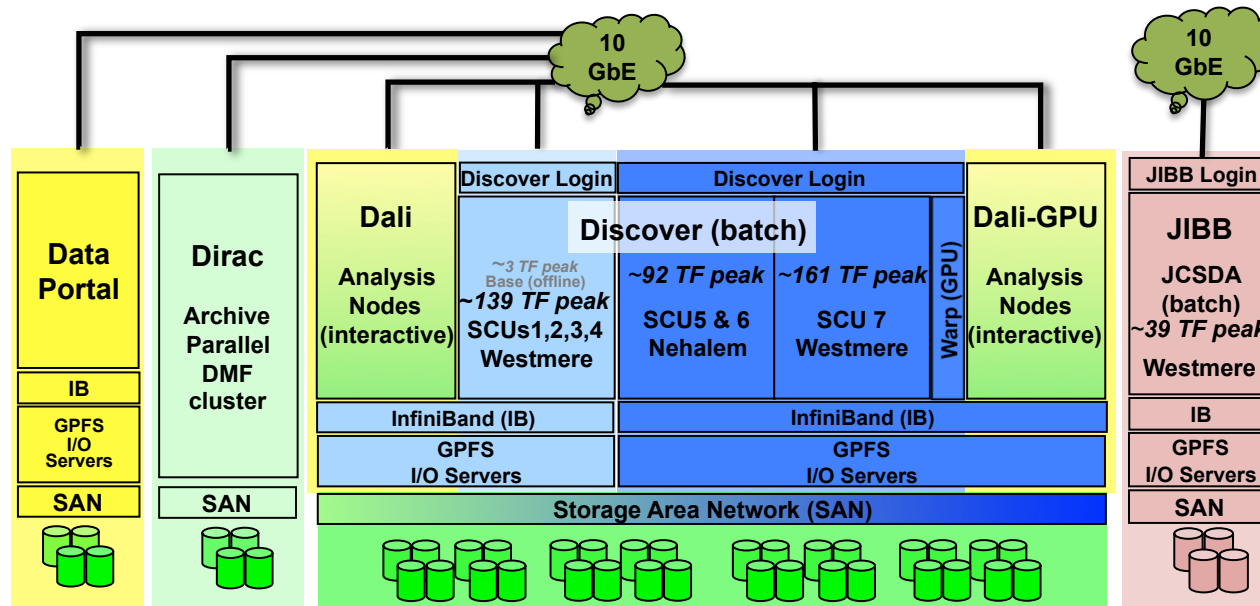
- Location of your files rooted in:
 - /archive
 - /discover
- Types of file systems
 - CXFS
 - GPFS
- Systems connectivity
 - Discover/Dali
 - Dirac

Ideal way of adding to archive

- Create files on `/discover/nobackup`
 - many small files
- Create tar on `/discover/nobackup`
 - one big file
- Move to `/archive`
 - one big file
- Reasons will become clear soon. Let's look at the environment

NCCS

- Discover/Dali - high performance computing cluster
- Dirac - archive system, mass storage
- JCSDA (*JIBB*) - NASA/NOAA collaboration
- Dataportal - Data sharing



Discover/Dali

- 35,560 cores
- 28,672 GPU streaming cores
- 3,394 nodes
- 400 TFLOPS
- 3.7 PB usable disk storage

Dirac

- 16 interrelated servers
- 30 PB archive holdings
- 960 TB usable disk space

JCSDA/JIBB

- 3,456 cores
- 288 nodes
- 39 TFLOPS
- 320 TB usable disk storage
- disk storage

Dataportal

- 16 HP blade servers
- 200 TB usable disk storage

NCCS without Data Portal and JIBB

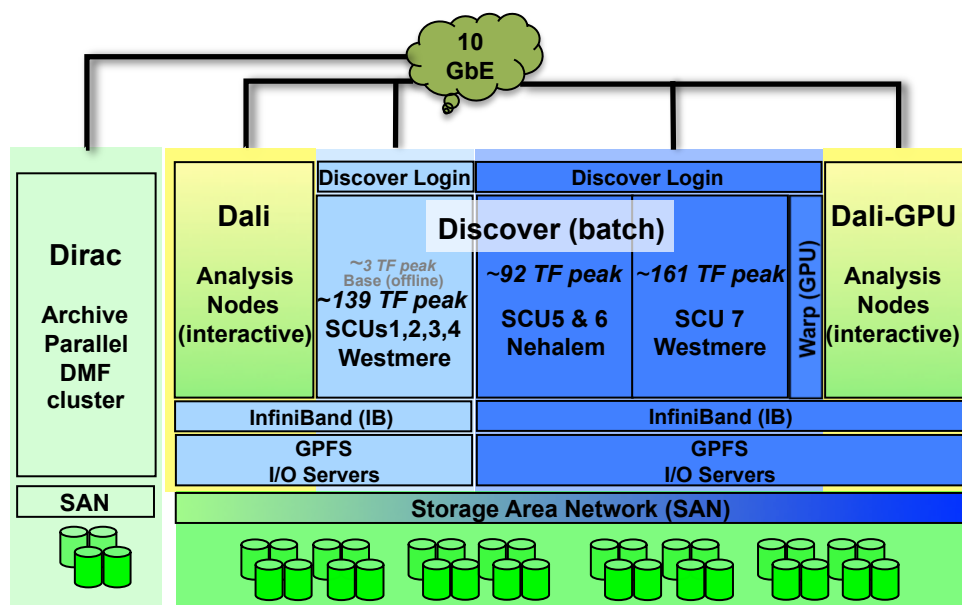
- Discover/Dali - high performance computing cluster
- Dirac - archive system, mass storage, DMF

Discover/Dali

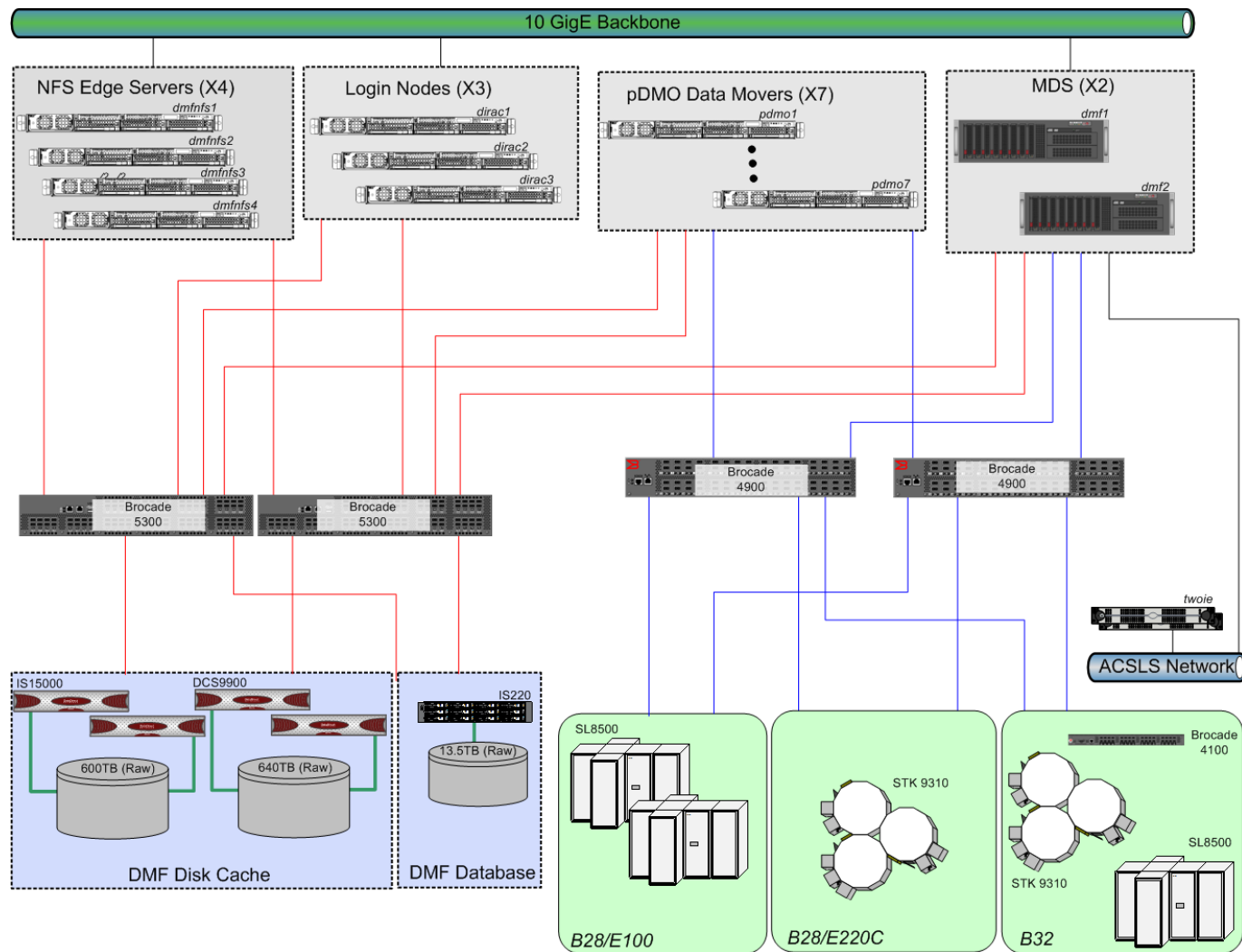
- 35,560 cores
- 28,672 GPU streaming cores
- 3,394 nodes
- 400 TFLOPS
- 3.7 PB usable disk storage

Dirac

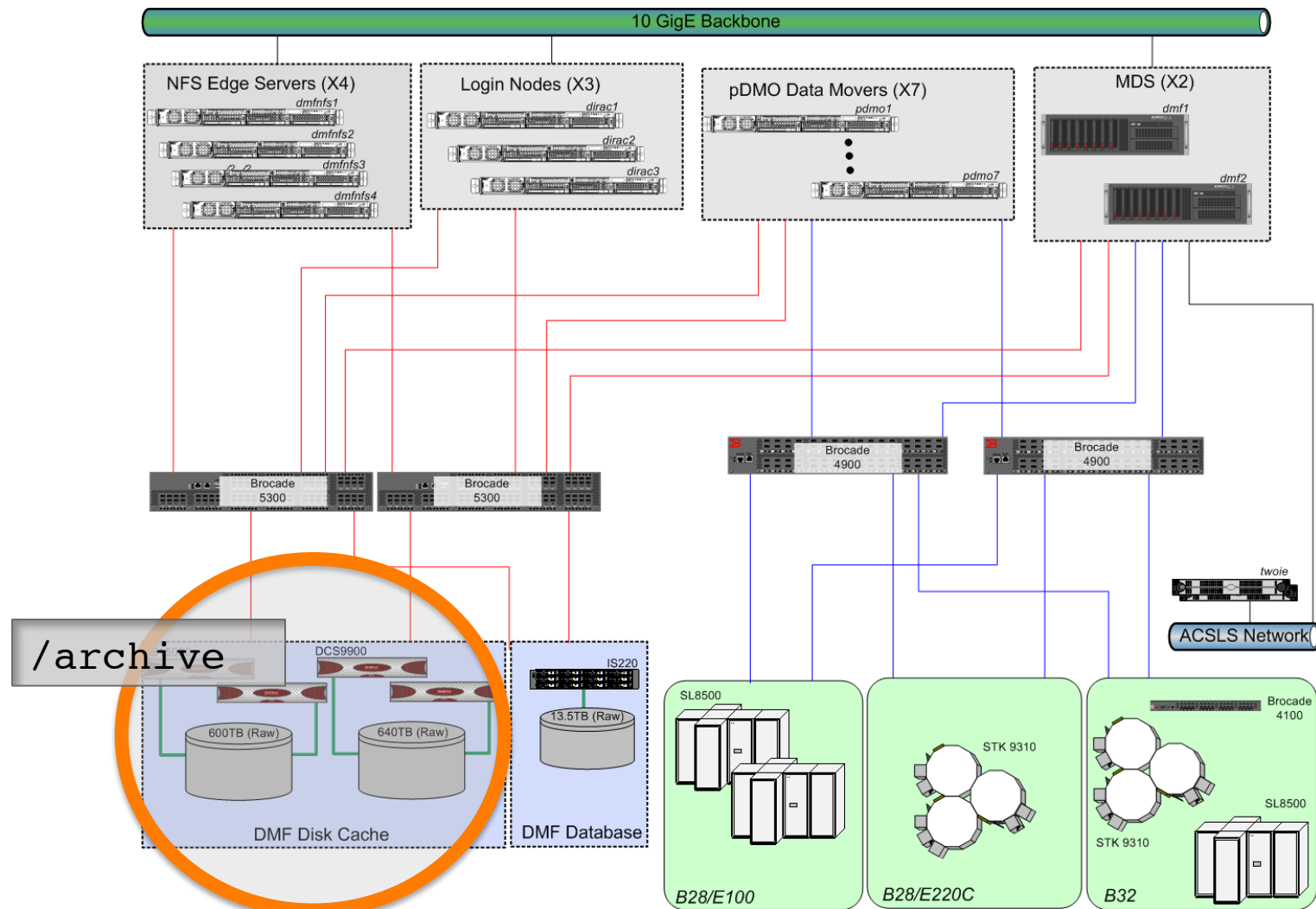
- 16 interrelated servers
- 30 PB archive holdings
- 960 TB usable disk space



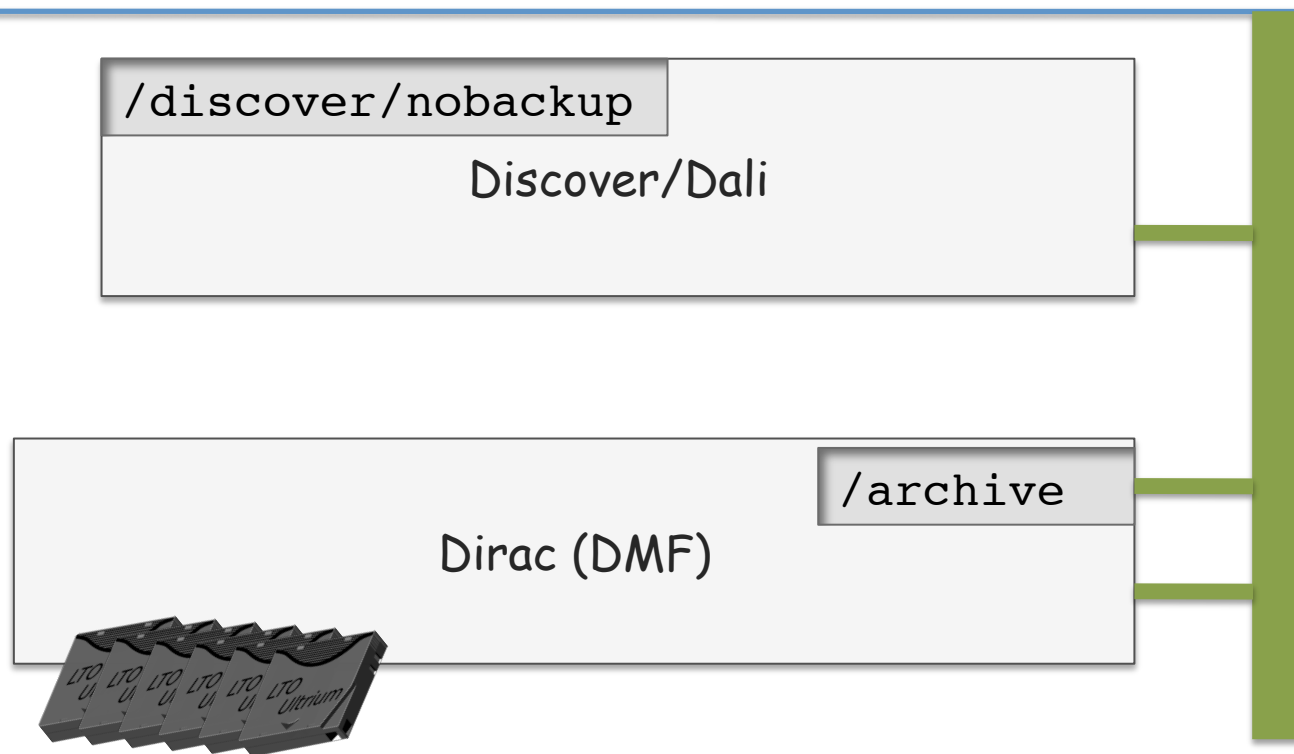
Data Migration Facility (DMF)



Data Migration Facility (DMF)

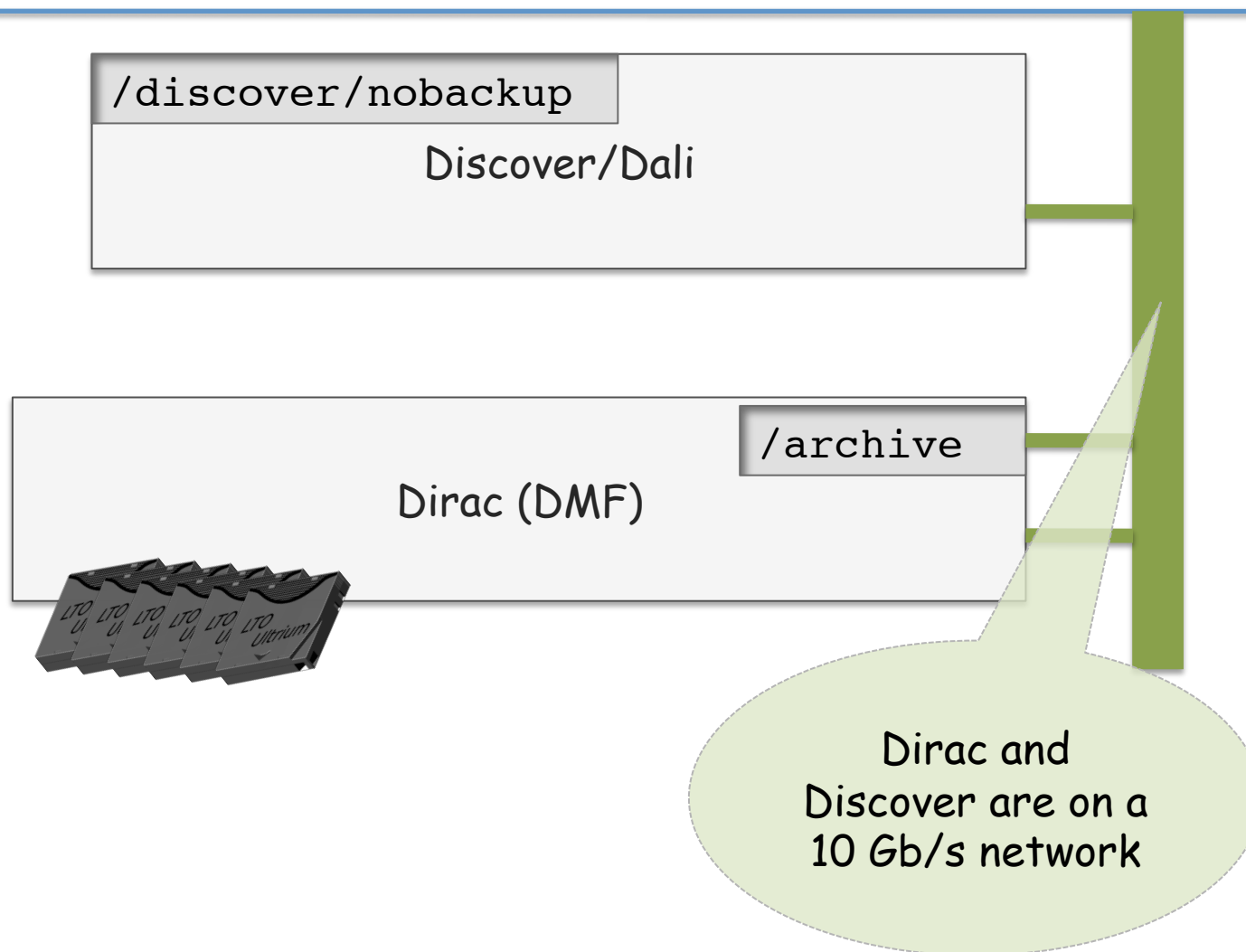


Connectivity simplified

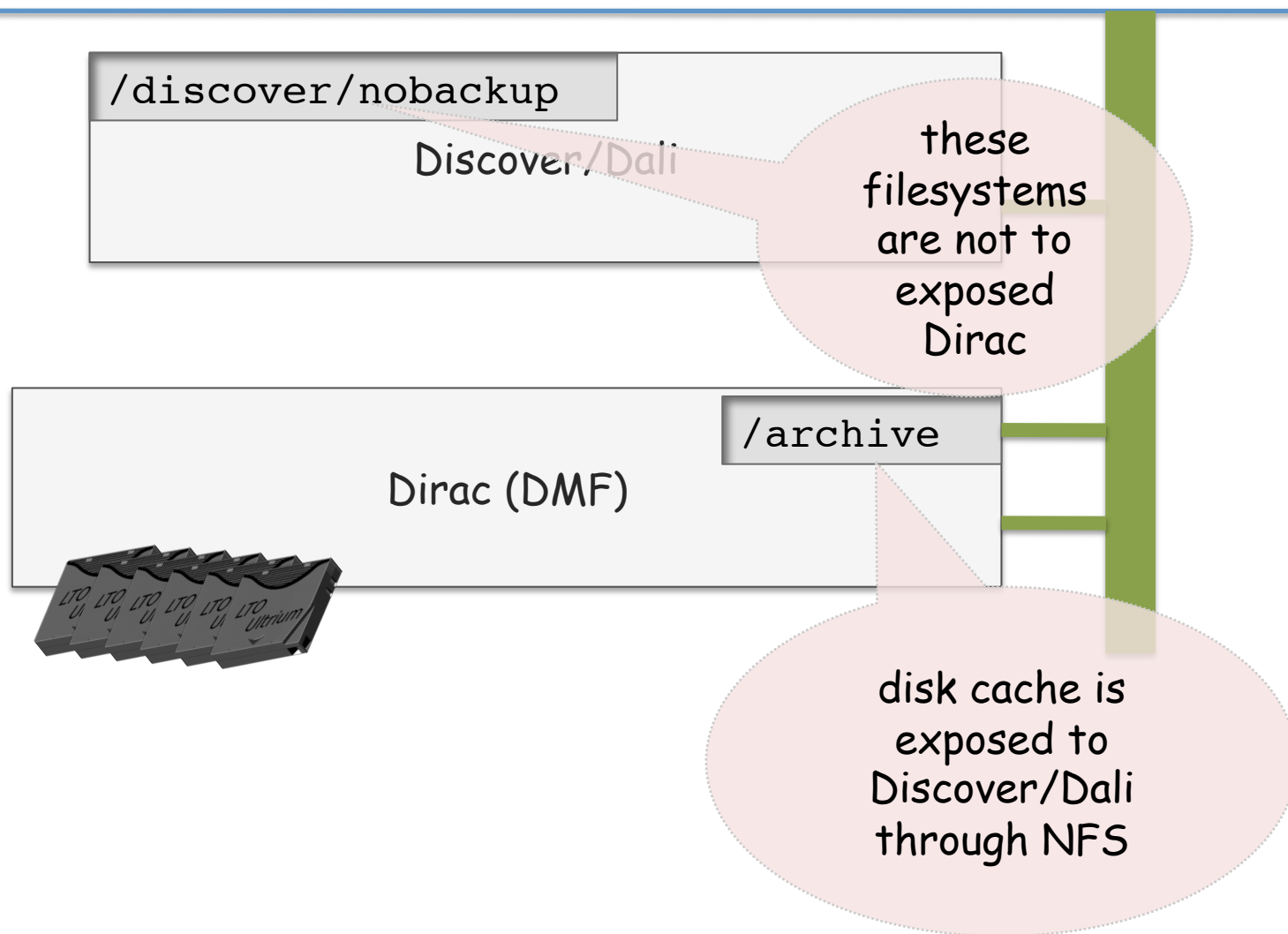


DMF: data migration facility

Connectivity simplified



Connectivity simplified



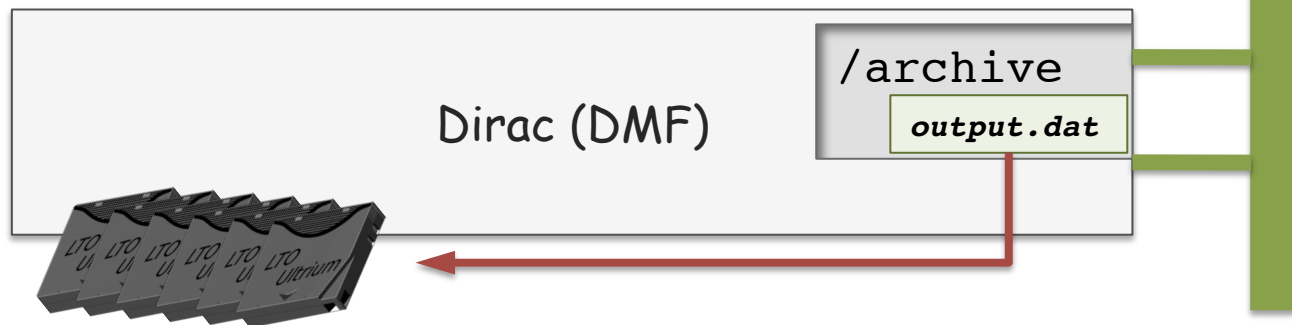
How does DMF work?

- 1) a file is written to disk cache --- /archive



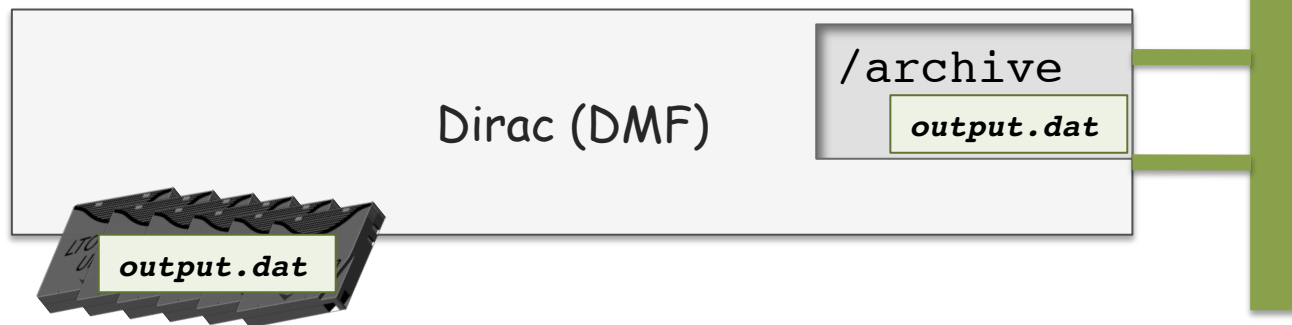
How does DMF work?

- 1) a file is written to disk cache --- /archive
- 2) DMF sees this and soon begins migrating to tape



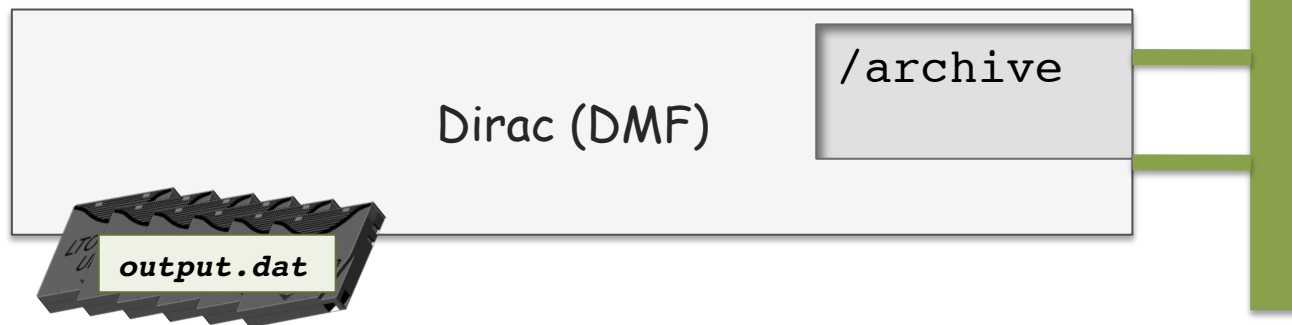
How does DMF work?

- 1) a file is written to disk cache --- /archive
- 2) DMF sees this and soon begins migrating to tape
- 3) soon there are two copies: tape and disk cache
(3 for files where two tape copies have been requested)

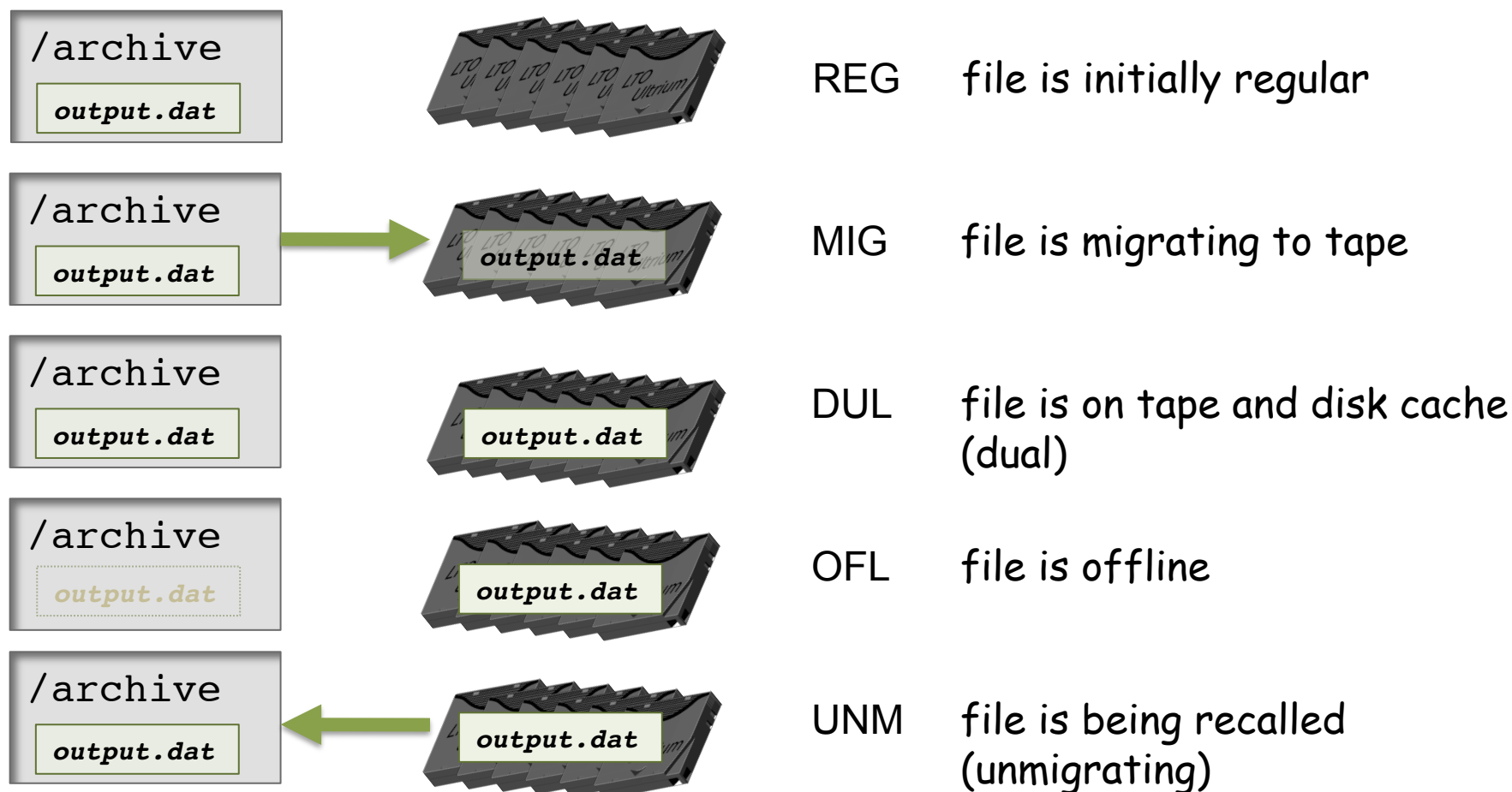


How does DMF work?

- 1) a file is written to disk cache --- /archive
- 2) DMF sees this and soon begins migrating to tape
- 3) soon there are two copies: tape and disk cache
(3 for files where two tape copies have been requested)
- 4) eventually file is on tape only



DMF states of a file as reported by dm1s



Questions about "taped" files

- How do I get it back from tape to disk cache
- How do I delete a file from disk cache
- How do I delete a file from tape
- How do I force a migration now
- What is the state of my file?
- What does my quota mean on Dirac?
- How do I turn a file I don't need right now into OFL from DUL? (for users wanting to be good citizens)
- Can I get two tape copies made
 - my data is very precious, I need more assurance



Answers

How do I get it back from tape to disk cache?

`dmget` command

See NOTE at the end

How do I delete a file from disk cache?

`rm` command - also removes access to the tape file

See NOTE at the end

How do I delete a file from tape?

`rm` command

space taken up by deleted files is reclaimed in a while (one or two weeks)

How do I force a migration now?

Normally DMF schedules migration, but the `dmpu` command will force an earlier migration

Answers

What is the state of my file?

`"dmls -l"` command


How do I turn a file I don't need right now into OFL from DUL?

`"dmput -r"` will release the disk blocks of a DUL file so that it becomes OFL. This can be run on a REG file also, but the disk blocks will not be released until the file has been written to tape.

`"dmput"` (with no `-r`) on a REG file will also initiate migration without immediate release of space on the disk cache. Its utility is in trying to keep your files spread across as few tapes as possible.

dmls and dmget

```
$ dmls -l $ARCHIVE/results
total 7601744
-rw----- 1 userid k3001 134217728 2011-07-28 16:35 (DUL) data.tar
-rw----- 1 userid k3001 536870912 2011-07-28 16:35 (OFL) data2.tar
```



dmls, dmget
also available from
Discover / Dali

```
$ dmget data2.tar
... wait for a while ...
$ dmls -l
total 7601744
-rw----- 1 userid k3001 134217728 2011-07-28 16:35 (DUL) data.tar
-rw----- 1 userid k3001 536870912 2011-07-28 16:35 (DUL) data2.tar
```

dmls and dmget

before

```
$ dmls -l data2.tar  
-rw----- 1 userid k3001 536870912 2011-07-28 16:35 (OFL) data2.tar
```

issue command

```
$ dmget data2.tar
```

during

```
$ dmls -l data2.tar  
-rw----- 1 userid k3001 536870912 2011-07-28 16:35 (UNM) data2.tar
```

after

```
$ dmls -l data2.tar  
-rw----- 1 userid k3001 536870912 2011-07-28 16:35 (DUL) data2.tar
```

dmls and dmget from Discover/Dali (caveats)

```
dali $ dmls -l  
dali $ dmget file1 file2..
```

passwordless ssh must be set up right
long pathnames are limited like in ssh

/discover/nobackup

Discover/Dali

/archive

Dirac (DMF)

many files?
long pathnames?
best to login on dirac

Answers (dmtag)

- Can I get two tape copies made?
 - NCCS makes single tape copies by default but you can get an extra copy if you ask for it
 - use the `dmtag` command
 - available on Discover, Dali, datamove queue
 - used for changing the *sitetag* of files in the archive
 - *sitetag* tells the system to make 2 copies

```
$ dmtag -t 2 list-of-archive-files
```

```
$ dmtag -t 2 < filelist.txt
```

also
accepts
list from
stdin

anything other than "2" will get
you a single copy

dmtag

check the sitetag of an archive file with **dmtag** (no arguments)

```
$ dmtag list-of-archive-files
```

```
$ dmtag test_*
0 /cx fsm/cache06/users/g05/username/test_t10kc.10GB
0 /cx fsm/cache06/users/g05/username/test_t10kc.2gb
0 /cx fsm/cache06/users/g05/username/test_t10kc.2gb.2

$ dmtag -t 2 test_*
$ dmtag test_*
2 /cx fsm/cache06/users/g05/username/test_t10kc.10GB
2 /cx fsm/cache06/users/g05/username/test_t10kc.2gb
2 /cx fsm/cache06/users/g05/username/test_t10kc.2gb.2
```


dmtag (caveats)

- Passwordless ssh must be set up (just like for `dmget`) when you run the command from Discover/Dali
<http://www.nccs.nasa.gov/primer/getstarted.html#passwordless>
- Inconsistency with symbolic links
 - changing the sitetag of the link in general changes the tag of the file to which the link refers
 - in some cases `dmtag` will report the sitetag of the link and not the file

Interaction between shell and dmtag

Scenario: there are two files only, `list` and `list.2`

C-shell (**csh** or **tcsh**) will not report errors about non-existent files!

```
% dmtag list* clap*  
0 /cx fsm/cache06/users/g05/username/list  
2 /cx fsm/cache06/users/g05/username/list.2
```

Bourne shell (**bash** or **ksh**) behaves better:

```
$ dmtag list* clap*  
clap* does not exist or the file it points to does not exist  
  
... more diagnostic messages ...  
  
0 /cx fsm/cache06/users/g05/username/list  
2 /cx fsm/cache06/users/g05/username/list.2
```

Impact of single tape copy

- Longer retrieve times for some files
- Some (few?) archive files will be unrecoverable

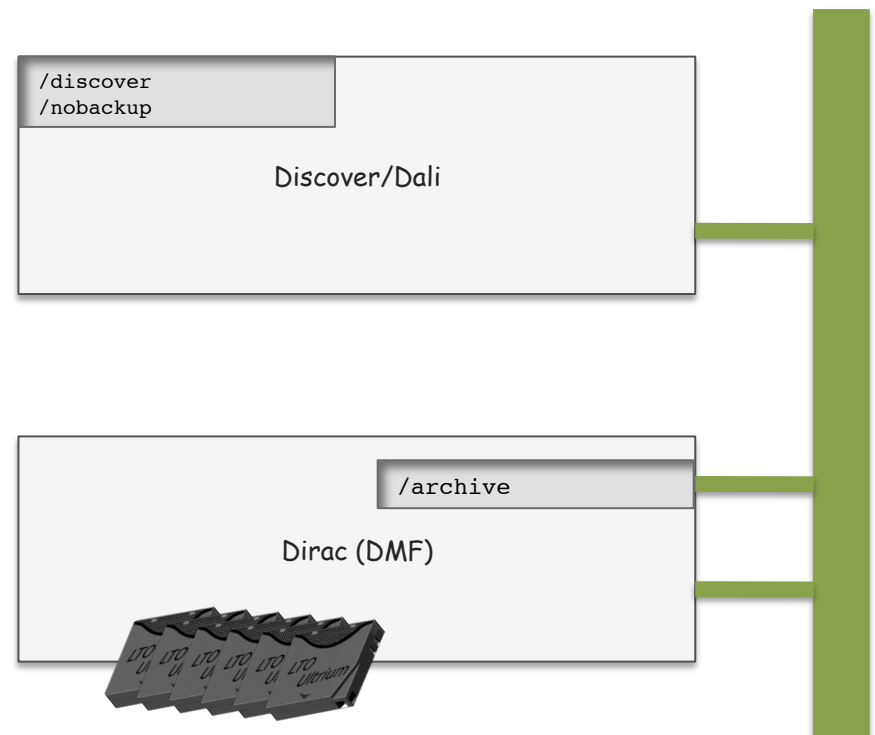
Ramifications

- logged in on Discover/Dali
 - slow access to `/archive`
 - fast access to `/discover/nobackup`, `/discover/home`
 - no access to `/archive/home`
- logged in on Dirac
 - fast access to `/archive`, `/archive/home`
 - no access to: `/discover/nobackup`, `/discover/home`
- So, where should I be logged in when moving data?

Q: What is wrong with this scenario?

```
dali $ cd /discover/nobackup/modeldir
dali $ simulation.out ./outputdir
dali $ cp ./outputdir/* /archive/outputdir/
dali $ tar cf /archive/bigdata.tar /archive/outputdir/*
```

- 1) simulation creates thousands of small files
- 2) copy all files to archive
- 3) create a tarball in archive



A1: What is wrong with this scenario

```
dali $ cd /discover/nobackup/modeldir
dali $ simulation.out ./outputdir
dali $ cp ./outputdir/* /archive/outputdir/
dali $ tar cf /archive/bigdata.tar /archive/outputdir/*
```

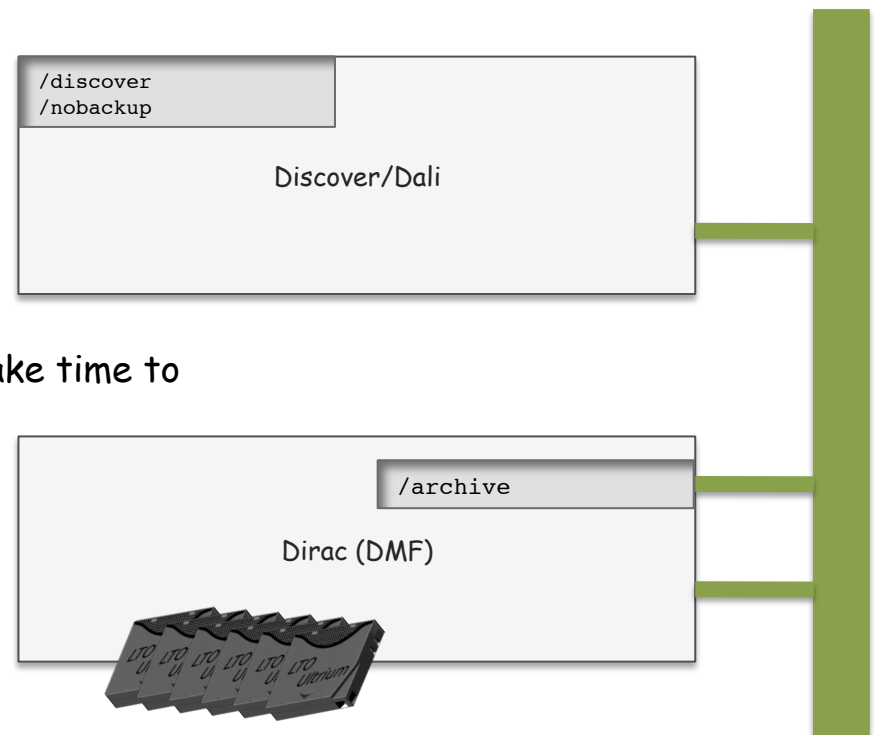
A1:

Thousands of files hit /archive...

Unless removed quickly, migration may begin

Unintended consequences:

- 1) multiple tapes may be affected
- 2) even if you delete files, the tape will take time to be recycled

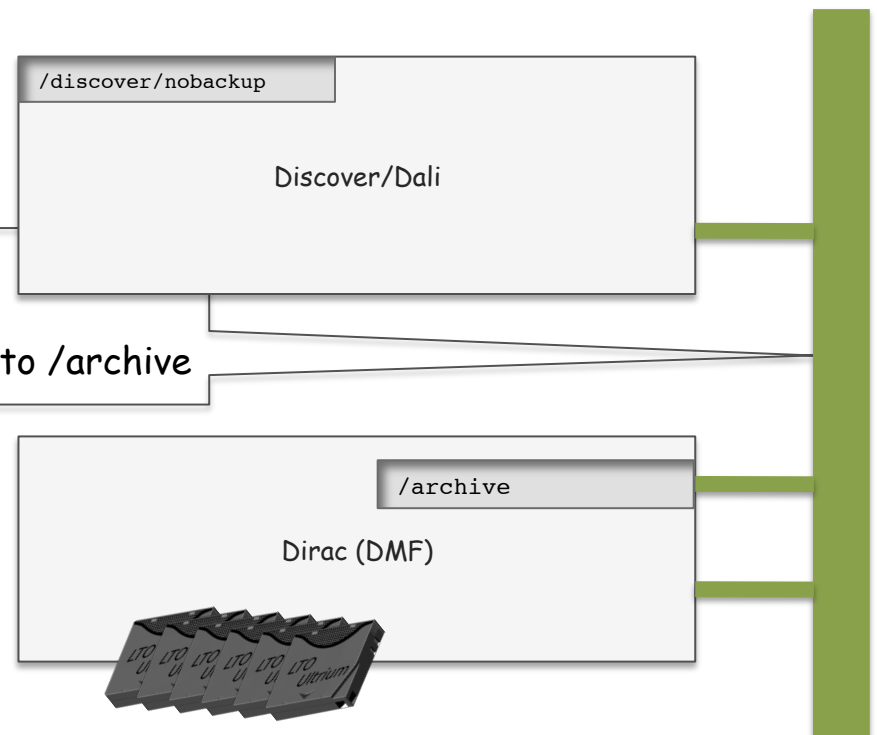


A2: What is wrong with this scenario

```
dali $ cd /discover/nobackup/modeldir
dali $ simulation.out ./outputdir
dali $ cp ./outputdir/* /archive/outputdir/
dali $ tar cf /archive/bigdata.tar /archive/outputdir/*
```

A2:

1. Dali reads files over the **network** from /archive
2. tar process on Dali creates the tarball
3. tarball is written over the **network** from Dali onto /archive

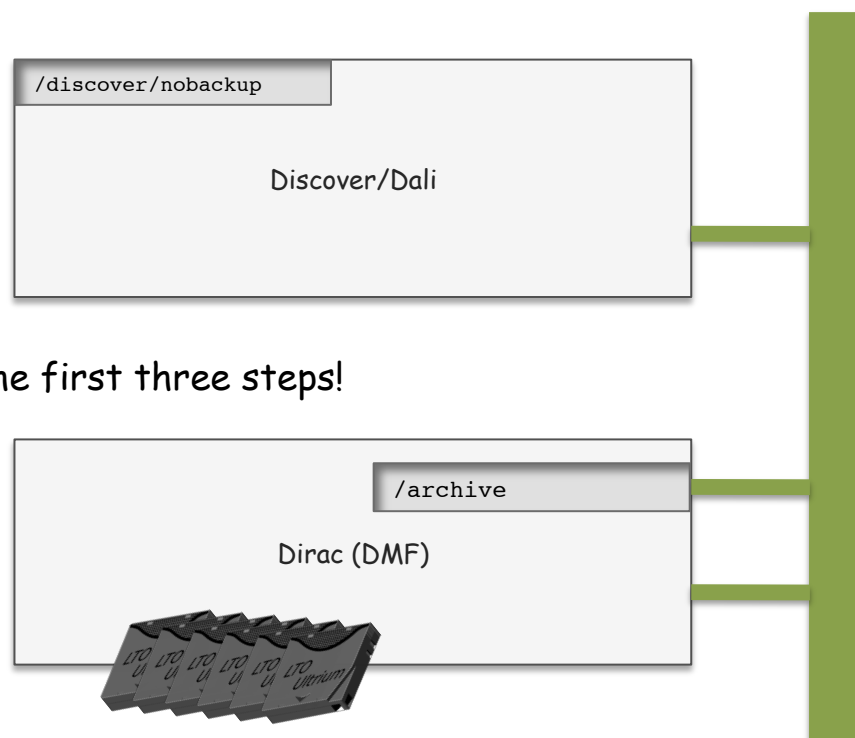


What is the remedy?

```
dali $ cd /discover/nobackup/modeldir  
dali $ simulation.exe ./outputdir  
dali $ tar cf /discover/nobackup/tardir/bigdata.tar ./outputdir/*  
dali $ cp /discover/nobackup/tardir/bigdata.tar /archive/outputdir/
```

but we recommend using
datamove nodes instead ...

No flow of any data to /archive during the first three steps!



Move large files with 'datamove' queue

interactive job

```
dali $ qsub -I -q datamove -l walltime=01:00:00
qsub: waiting for job 922008.borgpbs1 to start
qsub: job 922008.borgpbs1 ready
borg $ cp /discover/nobackup/tardir/bigdata.tar /archive/outputdir
```

Move large files with 'datamove' queue

as a batch job

```
#!/bin/bash
#PBS -S /bin/bash
#PBS -N mycopyjob
#PBS -l walltime=00:01:00
#PBS -j oe
#PBS -q datamove
#PBS -W group_list=g9999
source /usr/share/modules/init/bash
module purge
cp /discover/nobackup/tardir/bigdata.tar /archive/outputdir
```

Recap

- Do not
 - use /archive like a normal file system!
 - create, modify and delete small files
 - edit, compile, debug, ... edit
 - put anything there unless you want it moved to tape
 - run commands from Dali or Discover where both source and destination files are on the archive, because this needlessly uses up network bandwidth (e.g. tar , grep, wc, ...)
- Do
 - create large tar files somewhere else
 - copy large files to the disk cache using datamove nodes
 - if you want to unmigrate or remove more than a few hundred files on the archive system **PLEASE CALL OR EMAIL NCCS SUPPORT!**

301 286-9120 support@nccs.nasa.gov

(we can do things that user's can't)

THE END?

Acknowledgments

Nicko Acks
Fred Reitz
Ellen Salmon
Tom Schardt
Adina Tarshish

<http://www.nccs.nasa.gov/primer/data.html>

E-mail: support@nccs.nasa.gov

Tel: (301) 286-9120